Expr ss Mail Label N .: EL 979045045 US
Dat Mailed: N vemb r 25, 2003

UNITED STATES PATENT APPLICATION FOR GRANT OF LETTERS PATENT

PATRICK A. HOSEIN
INVENTOR(S)

QUEUING DELAY BASED RATE CONTROL

COATS & BENNETT, P.L.L.C.
P.O. Box 5
Raleigh, NC 27602
(919) 854-1844

QUEUING DELAY BASED RATE CONTROL

BACKGROUND OF THE INVENTION

[0001] The present invention generally relates to wireless communication networks, and particularly relates to rate control in such networks.

[0002] Rate control is a form of radio link adaptation wherein the transmission rate from a transmitter to a receiver is adjusted during ongoing communications responsive to changing signal quality, network loading constraints, etc. As an example, a wireless network base station may enforce common or per-user reverse link rate control to maintain reverse link loading at or around some targeted level.

[0003] Base stations also may enforce or otherwise supervise reverse link rate control of individual mobile stations to meet Quality-of-Service (QoS) requirements that specify maximum delay or jitter limits for reverse link transmissions for particular mobile stations. For example, a given mobile station may run one or more applications, each having its own logical "service instance," and each potentially having its own QoS requirements.

[0004] Supporting reverse link rate control in this context, the base station tracks or otherwise monitors indications of reverse link performance for each mobile station, so that it can determine when rate adjustments are required for the individual mobile stations to meet QoS or other requirements. Such monitoring requires the mobile stations to provide, i.e., transmit, reverse link information to the base station. For example, the mobile stations may provide the base station with information regarding their transmit buffer sizes as an indication of whether their reverse link rates should be adjusted upward or downward.

[0005] With this approach, for example, the base station may grant a higher reverse link rate to a mobile station that has more than a certain amount of pending transmit data buffered. In other words, a large amount of pending transmit data at the mobile station may trigger the base station to grant a higher rate for one or more subsequent transmit periods. Obviously, the base station can provide such control only when it is provided with transmit buffer information from the mobile

stations. Thus, the need for additional signaling between the mobile stations and the base station is one drawback of this approach.

[0006] Another drawback stems from the approach's failure to directly indicate a pending service problem, i.e., knowledge of a given mobile's transmit buffer size does not equate to direct knowledge of whether the mobile station's reverse link performance is at risk of violating QoS or other service constraints. For example, the average reverse link throughput of the mobile station may be quite high at the current time and, thus, one would expect even a relatively large transmit buffer to drain quickly. Thus, in addition to receiving transmit buffer size reports from the mobile stations, which undesirably adds overhead signaling to the finite-capacity reverse link, the base station generally has to monitor other conditions, or calculate additional metrics, to determine whether rate adjustments are needed for particular mobile stations.

SUMMARY OF THE INVENTION

[0007] The present invention comprises a method and apparatus wherein transmit rate adjustments are triggered or otherwise initiated based on expected transmit queuing delays. For example, if a first transceiver is transmitting data to a second transceiver subject to one or more service constraints, e.g., delay, jitter, etc., it may initiate rate changes for that radio link responsive to its evaluation of expected transmission queuing delays. For example, if the current queue size and current average throughput are such that length of time to empty the queue likely will violate a delay constraint, then the first transceiver may request or otherwise initiate a rate increase. In general, transceivers evaluate their expected queuing delay(s) in light of known service constraints and determine whether or not to initiate or request correspondingly appropriate rate changes.

[0008] In an exemplary embodiment, the present invention is applied to the reverse links between mobile stations and base stations in a wireless communication network, although it should be understood that the present invention can be applied to wireless, wired (electrical, optical, etc.) communication links. Thus, an exemplary method of reverse link rate control at a mobile station

comprises determining targeted queuing delays for reverse link transmit data, monitoring transmit data queue sizes and reverse link throughput at the mobile station, and generating reverse link rate requests based on the transmit data queue sizes, the reverse link throughput, and the targeted queuing delays. The mobile station may determine targeted delays based on QoS information received from the network that may be specific to each service instance being supported by the mobile station. Thus, the mobile station may receive targeted queuing delay information for one or more service instances being supported by the mobile station, periodically calculate an expected queuing delay for each service instance, and request a reverse link rate increase if any expected queuing delay exceeds a first delay value based on a targeted delay for the corresponding service instance, or request a reverse link rate decrease if the expected queuing delay for each service instance falls below a second delay value based on the targeted delay for the service instance.

[0009] Another exemplary embodiment comprises receiving targeted queuing delay information for one or more service instances being supported by the mobile station, and periodically calculating an overall data rate required to achieve targeted queuing delays for the service instances and requesting a rate change based on the overall data rate. The mobile station may periodically calculate an overall data rate required to achieve targeted queuing delays for the service instances by calculating a required data rate for each service instance needed to achieve the targeted queuing delay for that service instance in a next rate control period, and calculating the overall data rate based on the required data rates of the service instances. With the overall data rate thus calculated, the mobile station may request the nearest appropriate defined data rate, assuming that the mobile station is constrained to operate at one in a set of defined data rates.

[0010] Alternatively, the mobile station may use a rate dithering approach, wherein it maps the desired overall rate into an "effective" or "virtual" data rate that can be achieved using one or more combinations of defined data rates. The mobile station would thus request the appropriate virtual rate. In turn, an exemplary base station comprises transmitter circuits to transmit signals to a plurality of mobile stations, receiver circuits to receive signals from a plurality of mobile stations, and

processing circuits, including a rate control processor, to determine whether to deny or grant rate adjustment requests received from one or more mobile stations and to grant non-standard rate requests by mapping each non-standard rate request into a standard set of rates based on selecting one or more combinations of the standard rates.

[0011] Of course, the present invention is not limited by these exemplary details. Moreover, those skilled in the art will recognize additional features and advantages provided by the present invention upon reading the following discussion, and upon viewing the accompanying figures.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] Fig. 1 is a diagram of an exemplary wireless communication network according to an exemplary embodiment of the present invention.

Fig. 2 is a diagram of exemplary queuing delay based rate control processing.

Fig. 3 is a diagram of exemplary event-triggered rate control processing.

Fig. 4 is a diagram of exemplary base and mobile stations.

Fig. 5 is a diagram of exemplary periodic averaging/dithering rate control processing.

Fig. 6 is a diagram of exemplary virtual rate mapping.

DETAILED DESCRIPTION OF THE INVENTION

[0013] Fig. 1 illustrates an exemplary wireless communication network 10 that communicatively couples a plurality of mobile stations 12 to one or more Public Data Networks (PDNs) 14, such as the Internet. As illustrated, network 10 comprises a Radio Access Network (RAN) 16 that is coupled to the PDNs 14 a Packet Switched Core Network (PSCN) 18, which may comprise various Internet Protocol (IP) packet routers, gateways, and one or more authentication/authorization entities. Those skilled in the art should thus appreciate that network 10 as depicted is simplified for clarity and in actuality may include additional elements, such as Circuit Switched Core Network (CSCN) coupling RAN 16 to the Public Switched Telephone Network (PSTN). Further, it should be

noted that network 10 may be configured according to various network standards, such as, among others, IS-2000 or Wideband CDMA (WCDMA) standards.

[0014] Regardless, RAN 16 provides the wireless interface to the mobile stations 12 via forward and reverse radio links that may support per-mobile forward and reverse link traffic and control channels, one or more "broadcast" channels used for paging and common control, and one or more shared channels subject to scheduled use. In any case, an exemplary RAN 16 comprises one or more base stations, each comprising a Base Station Controller (BSC) 20 and one or more associated Radio Base Stations (RBSs) 22. BSC 20 may include packet data processing functionality for direct connection to PSCN 18, or may be coupled to PSCN 18 through a Packet Control Function (PCF) 24 or like entity.

[0015] In an exemplary embodiment as applied to reverse link rate control of the mobile stations 12, each mobile station 12 has a rate-controlled reverse link channel, e.g., a Reverse Link Packet Data Channel (R-PDCH) assigned to it, and the data rate of that channel can be adjusted upward or downward by the RBS 22 responsive to rate control request messages sent from the mobile station 12. Thus, according to Fig. 2, network 10 may send QoS information, e.g., targeted queuing delay information or other service constraint information, to the mobile station 12 from which it determines targeted reverse link transmit queuing delays (Step 100). That is, if a given service instance being supported by the mobile station has specific delay or jitter constraints associated with it, then there are limits on how long data to be transmitted from the mobile station 12 for that service instance can "wait" in the mobile station's transmit queue.

[0016] In operation, then, mobile station 12 monitors its transmit buffer size, i.e., its transmit data queue(s), and keeps track of its average reverse link throughput (Step 102). By maintaining a running average or other periodically updated estimate of its reverse link throughput, the mobile station 12 can estimate how long given data will remain in its transmit queue, or how long it will take to drain data that already is queued. Thus, mobile station 12 can generate link rate change requests based on the average throughput, the queue size(s), and the targeted queuing delays

(Step 104). In other words, the mobile station 12 generates rate change requests based on determining expected reverse link queuing delays and evaluating those expected delays relative to one or more service constraints. Those skilled in the art will appreciate that the same or similar logic could be applied by the RBS 22 or BSC 20 to the forward link, wherein expected queuing delays could be used to trigger forward link transmit rate adjustments, changes in forward link scheduling priorities, or both.

[0017] Fig. 3 illustrates an exemplary event-triggered embodiment wherein the mobile station 12 determines targeted queuing delays for each service instance being supported by it (Step 110). In operation, then, mobile station 12 monitors the transmit queue size and throughput for each service instance (Step 112), and periodically evaluates the expected queuing delay for each service instance (Step 114). If the expected queuing delay of any service instance exceeds the targeted delay for that service instance (Step 116), mobile station 12 may initiate a rate increase by sending a rate change request message to RBS 22. If none of the expected queuing delays exceed their corresponding target delays, mobile station 12 may check whether all of the expected delays fall below corresponding targeted delays (Step 122). If so, mobile station 12 may initiate a rate decrease by sending a rate change request message to RBS 22. Note that the comparison of expected-to-targeted queuing delays for rate increases may be different than for rate decreases. That is, the mobile station 12 may implement hysteresis or other rate change smoothing control by calculating different target delays for use in Steps 116 and 120.

[0018] In looking at target delay calculations in more detail, one may assume that each mobile station 12 is provided with either a maximum Radio Link Protocol (RLP) buffer delay requirement or a target delay value that should be maintained in order to reduce jitter at the initiation of each service instance. One also may assume that the user class, if any, of the mobile station 12 is known. Such information, if desired, may be used to determine the settings for certain rate control parameters. For example, if a maximum delay threshold of τ_{max} is required then one may define an associated target delay value as $\tau = \theta \cdot \tau_{max}$, wherein the value of θ depends on, for example, the

outage probability of the application, i.e., the probability that the delay exceeds the specified threshold τ_{max}, and the user class of the mobile station 12. An exemplary method attempts to maintain the average queuing delay at τ. If a target delay is specified for jitter guarantees, then mobile station 12 may be configured to set the target delay to that value.

[0019] Thus, let τ_i denote the target delay for service instance *i*. On a periodic basis, such as every T seconds, mobile station 12 computes a filtered estimate u_i in bits-per-second (bps) of the throughput so far achieved for service instance *i*. It also monitors the present size b_i in bits of the transmit queue for the service instance and estimates the expected queuing delay of a pending RLP frame to be transmitted as by $d_i = b_i / u_i$. If $d_i > \alpha\tau_i$ for any service instance *i*, then mobile station 12 transmits a rate increase request to RBS 22. Conversely, if $d_i < \beta\tau_i$ for all service instances, the mobile station 12 sends a rate decrease request to RBS 22. Otherwise, mobile station 12 may choose to maintain its present reverse link rate. The parameters $0 \le \alpha \le 1$ and $0 \le \beta \le 1$ may be determined by the user class of mobile station 12. For example, both parameters can be set closer to one as the user class preference increases. In the case of a jitter constrained application these parameters also may depend on the tightness of the jitter constraints. Naturally this approach can be generalized to multiple thresholds.

[0020] In any case, the rate request transmitted from the mobile station 12 will depend on the threshold region within which the present delay estimate lies. Before sending a rate increase request, the mobile station 12 first checks to see if its power headroom is large enough to support the higher rate. If not, the request is not sent. Additionally, one can enhance the above operations by taking into account the effect of a rate request command. For example, if a rate increase request is granted then in the next rate control period the mobile station's queue will drain twice as fast, assuming that a grant results in a doubling of data rate. Conversely, if a rate decrease is granted then the queue will drain half as fast, again assuming that a rate decrease halves the data rate. Such information can be used in making better decisions for α, β, T, and the buffer size(s).

For example, the values of α and β can be set based on the granularity of the defined rate change steps. Here, one notes that the mobile station's buffer capacities implicitly are assumed to be sufficient to support a queue equal to the product of the maximum rate and the maximum target delay. If not, then Adaptive Queue Management (AQM) techniques may be used.

[0021] Mobile stations 12 may be configured to support the above and other exemplary embodiments of the present invention. By way of non-limiting example, Fig. 4 illustrates exemplary mobile station and RBS configurations that may be used to support the present invention as applied to either or both forward and reverse link rate control. Mobile station 12 comprises an antenna assembly 30, a receiver circuit 32, a transmitter circuit 34, a baseband processor circuit 36, including transmit buffer memory 38 and a rate control circuit 40, and further comprises a system controller 42 and an associated user interface 44. RBS 22 comprises receive/transmit antenna elements 50, pooled receiver circuits 52 and associated reverse link processing circuits 54, pooled transmitter circuits 56 and associated forward link processing circuits 58, and interface and control circuits 60, including a rate/scheduling control circuit 62. Those skilled in the art should appreciate that these mobile station and RBS depictions represent exemplary functional arrangements and, because these entities typically comprise one or more microprocessor/digital signal processors, or collections of such processing resources, other functional arrangements may be used as needed or desired.

[0022] Regarding mobile station 12, rate control circuit 40, which may be implemented in hardware, software, or some combination thereof, functions as a rate controller that perform exemplary rate control processing in accordance with the present invention. Thus, it may track or otherwise have access to average throughput information for all service instances being supported by mobile station 12, and may monitor transmit buffer queue sizes as part of its rate control operations. Thus, it may monitor queue size and throughput information for each service instance being supported by the mobile station 12, and generate rate control requests for transmission to RBS 22. In turn, RBS 22, e.g., via rate/scheduling control circuit 62, may respond to those requests

accordingly. For example, if mobile station 12 transmits a rate increase request, RBS 22 and/or BSC 20 may grant or deny the request according to ongoing loading conditions or other criteria.

[0023] In another exemplary embodiment, mobile station 12 may be configured to embody the processing logic of Fig. 4, wherein queuing delay-based rate control is performed on a periodic basis. Such an approach may be used to maintain the average queuing delay at or near some specified value. With this approach, the mobile station may, for each service rate, compute the reverse link data rate that, if used in the next control interval, results in an expected delay of τ_i at the end of the control interval. In other words, mobile station 12 consistently attempts to change the rate so that the targeted queuing delay is achieved at the end of the control interval. Thus, the rate, λ_i, that should be requested for the next rate control interval is given by,

$$\lambda_i = \frac{u_i \tau_i - b_i}{T} + u_i,$$

which gives the desired rate for service instance *i*. Thus, the overall data rate required for the next rate control service interval is determined based on the desired rates of all service instances included in the evaluation, which typically comprises all service instances being supported by the mobile station 12. Thus, the overall rate, λ, that should be requested for the next rate control interval that will allow the mobile station 12 to achieve the targeted queuing delays for all service instances of concern is given as $\lambda = \sum_{i=1}^{i=N} \lambda_i$.

[0024] Commonly, however, only defined data rates may be requested by mobile station 12. That is, the network standards on which network 10 is configured, e.g., IS-2000, WCDMA, etc., may provide for a set of defined reverse link data rates. Thus, mobile station 12 may be configured to map the overall desired data rate into the set of defined data rates based on selecting the closest defined rate that will allow it to at least meet the targeted queuing delays over the next control interval. The selected rate, $\tilde{\lambda}$, is thus requested by the mobile station 12.

[0025] As an alternative to choosing a defined rate, the mobile station 12 and network 10 may be configured to adopt a "rate dithering" approach, wherein the mobile station 12 and RBS 22 or BSC 20 are configured with "virtual rate" tables that represent effective data rates that can be achieved by assigning one or more combinations of defined data rates. For example, suppose that the mobile station 12 determines an overall desired rate, λ, as the summation of the targeted rates, λ_i, for all service instances i. Rather than mapping the overall rate into the defined rate set, mobile station 12 computes or otherwise selects a virtual data rate, e.g., from configured data in its memory, and transmits a request to the RBS 22 that identifies that virtual rate. In turn, RBS 22 or BSC 20 processes that virtual rate by determining the appropriate combination of defined rates that will effect that virtual rate over some defined number of transmit intervals. Such mapping may be based on configured virtual rate tables held in memory at the RBS 22 that maps virtual rates to corresponding combinations of defined rates and transmit intervals.

[0026] Thus, in one embodiment, both mobile station 12 and RBS 22 (or BSC 20) contain tables that map combinations of N (non-unique) rates to the corresponding average rate. Assuming K supportable rates, then such mapping results in K^N virtual" rates. Note that the method may take into account re-transmissions, etc., in such computations. Mobile station 12 thus can map the desired overall rate to the nearest virtual rate and send the request to the RBS 22, which in turn looks up the sequence of defined rates that should be granted to the mobile station 12 for the next N frames that achieves this virtual rate. Use of virtual rates in this manner reduces rate fluctuations associated with servicing the transmit queues, while still meeting the desired QoS guarantees. By way of non-limiting example, assume that $N = 2$, and that defined rates 2, 3, and 4 are supported and an average of two transmissions is needed per frame. The possible virtual rates (after H-ARQ) are 1, 1.25, 1.5, 1.75, and 2. If, for example, the mobile station 12 requires a rate of 1.23 it can request a rate of 2 x 1:25 = 2:5 from the RBS 22. In turn, the RBS 22 can allow the mobile station 12 to achieve this rate by granting it a rate of 3 for a first transmit frame and a rate of 2 for the second transmit frame.

[0027] Fig. 5 illustrates exemplary processing logic to implement the above rate average and rate dithering methods. Processing begins with the assumption that mobile station 12 receives targeted queuing delay or other QoS information as needed for each service instance, and that it maintains ongoing throughput and queue size information values as needed for each service instance (Step 130). For periodic rate control adjustment, mobile station 12 may be configured to implement a rate control timer in hardware or software that sets the control interval for rate adjustments (Step 132). If it is time for a rate control adjustment (Step 134), mobile station 12 computes a target data rate needed to achieve the targeted queuing delay for each service instance i as explained above.

[0028] Mobile station 12 initializes i to 1 and sets its overall desired rate to "x," which may be zero or some other value (Step 136). For the ith service instance, mobile station 12 calculates the target rate required to achieve the targeted delay in the next control interval (Step 138), and adds that rate to the overall desired rate (Step 140). If there are more service instances (Step 142), mobile station 12 increments i and repeats the target rate and overall desired rate calculations for the next service instance.

[0029] Upon finishing such calculations for the last service instance, the mobile station 12 may perform either the defined rate or virtual rate mapping as described above (Step 146-1 or 146-2). That is, mobile station 12 may directly map the overall desired rate into the set of defined rates, or it may map that value into a set of virtual rates. In either case, mobile station 12 sends a rate request message for the next rate control interval (Step 148).

[0030] Assuming that mobile station 12 requested a virtual rate, Fig. 6 illustrates exemplary processing at RBS 22, wherein the RBS 22 receives the rate control request identifying the requested virtual rate (Step 160). RBS 22 accesses its stored virtual rate table(s) to look up the combination of defined rates and corresponding transmit intervals that should be used to achieve the virtual rate (Step 162). In this respect, rate/scheduling control circuit 62, or some other digital processing logic circuit in RBS 22, can be configured to access a memory-based lookup table that

is configured with the virtual-to-defined rate mappings. Regardless, RBS 22 determines the corresponding standard rates and grants them to mobile station 12 for the required number of frames (Step 164).

[0031] Of course, those skilled in the art will recognize that rate dithering using such virtual rates, or rate average as described above, are not essential to the present invention. Indeed, whether event-triggered, or driven by periodic rate averaging/dithering, the present invention provides a method whereby radio link rate adjustments are made based on the evaluation of expected/targeted queuing delays relative to QoS requirements or other performance constraints.

[0032] Such radio link adjustments, as noted earlier herein, are not limited to the reverse link and, indeed, the present invention provides exemplary queuing-based forward link control in one or more base station embodiments. For example, various third generation (3G) network standards, such as 1XEV-DV, 1XEV-DO, and WCDMA, use a shared high-speed channel in the forward link to transmit data for a plurality of data connections corresponding to a group of users, i.e., a given group of mobile stations 12. At any given instant, the shared channel typically carries traffic for only one data connection, but over time all users receive data via the shared channel based on "scheduling" operations carried out at RBS 22, wherein the RBS 22 transmits traffic for selected data connection(s) in each of an ongoing sequence of scheduling intervals.

[0033] Thus, in such embodiments, RBS 22 allows users to time-share one or more forward link communication channels, and assigns a scheduling utility function to each data connection of a user sharing the channel. In an exemplary embodiment, rate/scheduling control circuit 62 is configured to implement a scheduler that tries to maximize the total utility function, which may be based on joint evaluation of the individual utility functions. In any case, if a particular data connection has a delay constraint (e.g, a limit on the maximum delay or a jitter constraint) then this constraint can be included in the corresponding utility function. Thus, the scheduling priority of a given data connection may be made dependent on the expected queuing delays for that connection

and, in particular, may be made to depend on the expected queuing delays relative to that connection's targeted queuing delays.

[0034] For example, rate/scheduling control circuit 62 may receive or have access to QoS constraints associated with one or more of the data connections. Thus, a given data connection may have a jitter-imposed targeted queuing delay of qt, and the rate/scheduling control circuit 62 may be configured to maintain the queuing delay as close as possible to qt. As such, the utility function for that data connection can be configured as $U(q) = -(q - qt)^2$, where the expected queuing delay, q, is estimated as described above for the reverse link. RBS 22 maintains an estimate of the forward link throughput for the data connection and, whenever a scheduling decision is to be made, it estimates the delay as the ratio of the queue length and the throughput. Thus, RBS 22 can detect when the expected queuing delays of any data connection exceed that connection's targeted delays and adjust the scheduling priorities accordingly. Therefore, RBS 22 applies essentially the same logic as the mobile station does in making rate control requests based on queuing delays but instead of making rate requests, RBS 22 changes user scheduling priorities based on queuing delays.

[0035] Of course, it should be understood that rate/scheduling control circuit 62 can be configured to make scheduling adjustments and/or make forward link rate adjustments based on expected queuing delays. For example, if RBS 22 serves a plurality of mobile stations 12 using dedicated, rate-adjustable forward link communication channels, rate/scheduling control circuit 62 can be configured to initiate rate adjustments on those channels as a function of expected versus targeted queuing delays for outgoing traffic.

[0036] Therefore, the present invention may be applied to forward and reverse link rate control, forward link scheduling control, or to combinations thereof in a variety of wireless communication network types. Indeed, the present invention is not limited to wireless applications and may be applied to both wireless and wired communication links. As such, the present invention is not

limited by the foregoing discussion and, indeed, is limited only by the following claims and their reasonable equivalents.